NEW CONTENT-BASED FEATURES FOR THE DISTINCTION OF VIOLENT VIDEOS AND MARTIAL ARTS

Markus Hörhan and Horst Eidenberger

Vienna University of Technology
Institute of Software Technology and Interactive Systems
Favoritenstrasse 9/1882, 1040 Vienna, Austria
{first.last}@tuwien.ac.at

ABSTRACT

Real violence is unwanted content in video portals as it is forensically relevant in video surveillance systems. Naturally, both domains have to deal with mass data which makes the detection of violence by hand an impossible task. We introduce one component of a system for automated violence detection from video content: the differentiation of real violence and martial arts videos. In particular, we introduce two new feature transformations for jitter detection and local interest point detection with Gestalt laws. Descriptions are classified in a two-step machine learning process. The experimental results are highly encouraging: the novel features perform exceptionally well and the classification process delivers practically acceptable recall and precision.

Index Terms— Violence detection, content-based video analysis, local interest point detection, jitter detection, SVM classification.

1. INTRODUCTION

This paper describes a content-based solution for the detection of violent video content. More specifically, we focus on one step in a greater plan: the differentiation of user-generated violent videos (which are often objectionable content on video sharing websites) from martial arts videos (which are not). The novel methods are two general-purpose feature transformations and a categorization scheme that balances over- and underfitting.

The greater plan is a three step process for violence detection: First, retrieving everything from a source that is potentially violent, secondly, filtering out martial arts and similar content, and thirdly, classifying the remaining videos as violent or not. Potential applications include the forensic analysis of video surveillance content and automated blocking of unwanted content on video sharing websites. Both applications are of highest relevance today: For example, in the Vienna underground approx. 480MB video content is produced per second. Currently, without knowing the exact train/station number and time, it is not possible to retrieve forensically potentially relevant content. An automated process with fair precision would improve this situation drastically.

In many works on violence detection, the classification task is applied on highly discriminative film genres, e.g. horror and romance. In contrast, we propose a method, which automatically distinguishes between user-generated violent videos and martial arts videos. Obviously, this problem requires a more sophisticated approach since a lot of content similarities do exist between these genres. For this end, we present the first implementation of a novel feature for video genre identification which is based on high-level perceptual Gestalt principles. Theoretically, this feature was first described in [1]. In the remainder of the paper, we describe the novel feature transformations, the categorization approach, the ground truth and the experimental results. In addition, the next section summarizes relevant related work.

2. RELATED WORK

Only a few works about content-based violence detection can be found in the literature so far. In [2] a combination of two individual kNN classifiers (one for audio features and one for visual features) is used to distinguish between violence and non-violence video segments. The approach of [3] utilizes motion, blood detection, face detection and some film production rules together with an SVM to detect violence in movies.

The second research direction that is related to our work is video genre classification. For instance, in [4] semantic features and text features derived from the title, tags, and video description are used to perform web video genre classification. In the work of [5] average shot length, color variance, motion, lighting key and visual effects are combined to categorize action, drama and thriller films.

3. PROPOSED APPROACH

Below, we desribe the employed content-based features, classifiers and the ground truth. The test videos were segmented

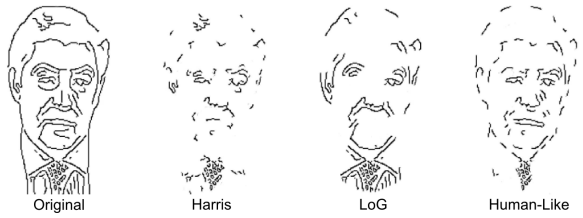

Fig. 1. The left edge map of a face is represented by points from a Harris corner detector and a Laplacian of Gaussian (LoG) operator. Many of the important face features (e.g. the eyes) are thereby lost. The rightmost image shows the GIP description. It preserves the perceptual features of the original stimulus well but does not produce a longer description than the LoG operator. [1]

automatically into shots using the free tool Shotdetect [6]. Shots are decomposed into frames, of which the novel *Gestalt Interest Points (GIP)*, jitter descriptions, color and SIFT are extracted. For extracting GIP, color and SIFT features, each 25th frame of a shot serves as a key frame; for jitter detection, the first n frames of each shot are taken as key frames. A detailed description of the descriptions and their semantic interpretation by categorization follows.

3.1. Gestalt Interest Points

To describe the local information in an image, we use the novel GIP and SIFT. GIP is based on the Gestalt law of closure [7] and the idea that, unlike other local methods, some weaker points are also useful as interest points in addition to the local extrema.

The Gestalt law of closure states that the perception of individuals fills in visual gaps in incomplete shapes. For example, humans are able to recognize a whole circle, even if there are gaps in its contour. For our approach this means that due to the Gestalt law of closure it is still possible to recognize what an image depicts, only by considering its local representation. This effect is shown in Figure 1. Obviously, such interest point sets are more useful for media understanding than points from which humans cannot identify the semantic content of an image. If the user cannot reconstruct the object from the interest points, how should the machine?

We developed several different methods for the implementation of GIP. It turned out that the following procedure leads to the best classification results for capturing body movement: Every n-th frame is converted to a grey scale image and then convolved with an edge operator (e.g. Sobel) to get the gradient vectors and gradient vector magnitudes for each image location. The resulting gradient image is split into $m - by - n$ (e.g. 10x10) macroblocks. For each block, we identify the three largest gradient magnitudes. These magnitudes are used to construct the image descriptor. It is composed of the three positions, the three orientations and the average of the three above selected magnitudes of each image block. Among the advantages of this straightforward scale-less implementation are the guarantee that the visual object shape is preserved in the description, and that heaps of high-curvature interest points are avoided: compared to SIFT, SURF and related methods the local description is more evenly distributed over the entire input signal without ending up in a global description.

3.2. Jitter

User generated videos often contain a considerable amount of jitter because such videos are often captured with handheld cameras. In our work, jitter is a very useful descriptor for user-generated video identification. To extract the jitter descriptors we use an adapted version of Matlab's video stabilization algorithm. For our purpose, we employ only the first part of this algorithm. To accomplish jitter detection, only the first 25 frames of each shot are taken into account. Experiments showed that this magic number represents a well-performing tradeoff between accuracy and speed in the given domain. The following procedure is applied to the key frames.

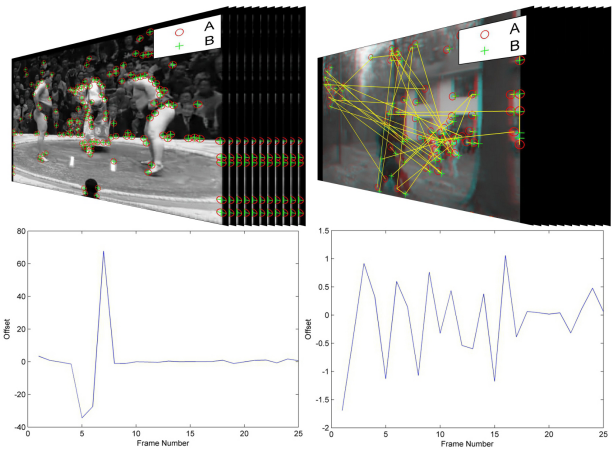

Fig. 2. *Left-Top*: Interest point correspondences between consecutive frames of a martial arts video. *Right-Top*: Interest point correspondences between consecutive frames of a violent video. *Left-Bottom*: Interest point displacements over 25 frames for a martial arts video. *Right-Bottom*: Interest point displacements over 25 frames for a violent video.

For the extraction of the jitter descriptor the displacements of selected interest points from a frame A to its successor B are measured. In both frames a corner detection algorithm selects interest points around salient image regions such as corners. In the next step, the selected points in frame A and frame B are matched using correspondences in their neighbor-

hood. Figure 2 shows the matched points with circles indicating points of frame A and plus symbols indicating points of frame B. The lines connecting points of frame A with points of frame B in the right top of Figure 2 represent the displacement vectors. In the example, they indicate a video that contains a significant amount of jitter. For performance reasons, we compute not more than 150 displacement vectors between each pair of key frames. The 150 x- and 150 y-components of the vectors are concatenated to form a 300 dimensional vector. Additionally, the mean is taken from these 300 values. For the 25 shot key frames, we get 25 mean values. Arranged in the temporal order of the frames they form a signal; cf. bottom of Figure 2. Means and variances of the interval lengths between neighboring zero crossings are then calculated from this signal. Our experiments show the value of this description type in the sketched domain.

3.3. Color

Color is a fast and effective way for describing visual media. We assume that color characteristics differ significantly between martial arts videos and user generated violent videos due to quality differences in the recording devices and the fact that professional videos undergo color correction in post-production. Therefore, the video quality of consumer videos should not come up to professionally produced content. Our color feature serves as a baseline for jitter descriptions. We utilize the HSV color space to describe each key frame. The feature vector is composed of mean and variance of each of the three HSV color channels. In early experiments we also investigated RGB and CIE XY color space, but the results indicated that the HSV color space performs better for the given task.

3.4. SIFT

Inspired by [8], we use SIFT combined with the Bag of Features (BoF) approach for our prototype. SIFT descriptors are extracted per key frame and transformed into histograms. We employ the SIFT-BoF implementation of the freely available CORI framework [9] to compute this feature that serves as a baseline for the GIP features in the prototype.

3.5. Classification

Our goal is to group the shots of the test videos into *two classes*: *martial arts shots* and *user generated violent shots*. To accomplish the task we use a separate Support Vector Machine (SVM) [10] for each of the four features and finally fuse the classifier decisions by applying a combination of a decision rule and majority voting. Combining SVM categorization together with rule-based method generalizes well for the given domain, because the SVM is a rigid method that avoids overfitting on the feature data. Feeding the SVM output into a decision rule, however, opens a new degree of freedom which allows to adapt to the semantics in the ground truth. The classification algorithm was composed based on experiments in Weka and is designed as follows.

Jitter detection plays a fundamental role in the classification process. If the jitter-SVM judges a shot with confidence score above an empirically defined threshold as a user-generated violent shot, the other classifiers are not considered anymore. The confidence score is computed from the relative number of frames of a shot classified as user generated violence. The judgments of the classifiers for the other features are taken into account if the confidence score of jitter detection is below the confidence threshold. In this case the final decision is derived by majority voting of the three non-jitter classifiers.

	Martial arts			User gen. violence		
	Rec.	Prec.	F1	Rec.	Prec.	F1
Jitter-based classification	51	77	61	95	85	90
Color-based classification	72	41	52	64	87	74
GIP-based classification	81	64	72	85	93	89
SIFT-based classification	81	44	57	65	91	76
Fused classification	83	96	89	99	94	97

Table 1. Recall, precision and f1-score of the proposed method for both video categories.

4. RESULTS

4.1. Dataset

We assembled a dataset that contains videos of the dataset from [8] and videos taken from youtube.com and gorillafights.com. The dataset is composed of 214 videos (10 hours in total), 107 for each category. The martial arts category consists of wrestling, sumo, boxing, kick boxing, karate and cache fight videos; the user-generated violent part is composed of indoor and outdoor videos that show people who fight against each other. In order to generate the ground truth for the dataset, we labeled each video by hand. As mentioned above, some videos of one category are very similar to videos of the other and, therefore, they can easily be misclassified.

4.2. Evaluation

For the evaluation of the proposed methods, the dataset was divided into two parts: one-third for training and the rest for testing. In the first step of the experiment, the categoriza-

tion process was conducted for each of the four features separately. Finally, the algorithm for fusing the classification outputs, as described in Section 3.5, was applied. Table 1 presents the results.

The f1-scores of the color-based classification are the lowest amongst all features. Nevertheless, this proves that color and also lighting conditions are in average to some degree different between the two categories. These differences can also be observed by the naked eye. An advantage of color is the low computational effort required for feature extraction, which is the lowest of all four considered features. The achieved f1-scores of the jitter are significantly higher than the f1-scores of the color feature. This justifies the statement above, that color (and SIFT) serves as baseline features for the two new approaches. Some scenes in martial arts videos are produced with handheld cameras and therefore the occurrence of jitter is not limited to user-generated violent videos. This leads to confusion and errors in the classification process. Some intended camera movements may also be misinterpreted as jitter.

The evaluation results for SIFT-based classification are slightly better than color-based classification. However, the GIP-based categorization clearly outperforms SIFT. This is a remarkable result, since both features are general-purpose methods applied on the same domain and ground truth. The result supports our thesis that weaker interest points are also useful features and not only the local extrema: Sometimes it makes sense to give up stronger points for isolated weaker ones that satisfy the Gestalt rules.

As we can see from the experimental results, we proved that the GIP are an effective feature in classification of video genres. Besides, jitter is very useful to discriminate user generated videos from professionally produced videos. The last row of Table 1 makes clear that all four described features in combination with the proposed classification algorithm are highly effective for distinguishing between martial arts videos and user generated violent videos.

5. CONCLUSIONS AND FUTURE WORK

Violent videos are often objectionable content on video-sharing websites but martial arts videos are not. Motivated by this problem, we developed a system which automatically discriminates between these video genres. For this task, we proposed the novel GIP feature and a novel method for jitter detection. Experiments showed that these features together with Color information and SIFT are well suited for the given task. Classification is performed using one SVM for each description type. A combination of a decision rule and majority voting fuses the classifier results.

We are currently working on a violence detection system that integrates the proposed algorithm. In a preprocessing step the system will separate violent videos from other video types. The resulting set of violent videos is processed to filter out martial arts videos. Furthermore, since the results have been exceptionally good, we will investigate the GIP feature and jitter detection in greater detail and for other domains. We are positive that both offer great potential for solving various problems of content-based retrieval.

6. REFERENCES

[1] Horst Eidenberger, *Fundamental Media Understanding*, atpress, Wien, 2011.

[2] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis, "Audio-visual fusion for detecting violent scenes in videos," in *Artificial Intelligence: Theories, Models and Applications*, S. Konstantopoulos et al., Ed. 2010, vol. 6040 of *Lecture Notes in Computer Science*, pp. 91–100, Springer Berlin / Heidelberg.

[3] Liang-Hua Chen, Hsi-Wen Hsu, Li-Yun Wang, and Chih-Wen Su, "Violence detection in movies," in *Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference on*, aug. 2011, pp. 119 –124.

[4] Linjun Yang, Jiemin Liu, Xiaokang Yang, and Xian-Sheng Hua, "Multi-modality web video categorization," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*, New York, NY, USA, 2007, MIR '07, pp. 265–274, ACM.

[5] Hui-Yu Huang, Weir-Sheng Shih, and Wen-Hsing Hsu, "Movie classification using visual effect features," in *Signal Processing Systems, 2007 IEEE Workshop on*, oct. 2007, pp. 295 –300.

[6] *Shotdetect*, http://shotdetect.nonutc.fr/ visited on November 8th 2012.

[7] K. Koffka, *Principles of Gestalt Psychology*, Lund Humphries / London, 1935.

[8] Fillipe D. M. de Souza, Guillermo C. Chavez, Eduardo A. do Valle Jr., and Arnaldo de A. Araujo, "Violence detection in video using spatio-temporal features," in *Proceedings of the 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images*, Washington, DC, USA, 2010, SIBGRAPI '10, pp. 224–230, IEEE Computer Society.

[9] R. Sorschag, "Cori: A configurable object recognition infrastructure," in *Signal and Image Processing Applications (ICSIPA), 2011 IEEE International Conference on*, nov. 2011, pp. 138 –143.

[10] B. Schlkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press / Cambridge, MA, 2002.